

# Why the word should become the central unit of phonetic speech research

Hans G. Tillmann  
*IPSK, Universität München*

The purpose of this paper is

- (i) to assert that phonetics as speech science needs a fundamentally new theoretical orientation (leading also to a quite new research agenda);
- (ii) to argue that this can be most easily<sup>1</sup> achieved if the word which to all native and non-native speakers of a language is the most naturally given unit of speech should also be considered the most central unit of speech science;
- (iii) to promote the idea that phonetics - taken in this way and in close connection with the other basic speech science, semantics - will become a much more useful discipline in the context of the development of speech technologies for the coming so-called information society; and finally
- (iv) to present along with earlier ideas published by the present author a rather complete picture of the phonetic processes which are basic in the production of natural speech acts.

My personal impression concerning the present situation of speech technology and spoken language processing is that despite all the remarkable successes of HMM- and NN-applications, a local minimum has been reached. For the further development of speech technology it will not be enough just to collect still more data for the training and testing of SLP-systems. Much more money has to be invested into new speech research and further development of theoretical concepts is needed. Therefore, and besides introducing the word as the central unit of phonetic speech research, I will also explicitly reintroduce the classical experimental phonetic concept of *systematic modification* (as apposed to random statistical variation of the phonetic forms of utterances<sup>2</sup>).

---

<sup>1</sup> In this paper the terms 'easy' and 'simple' are used with the meaning of Herrmann Paul's (1898:4) statement "dass ich nur für diejenigen schreibe, die mit mir der Überzeugung sind, dass die Wissenschaft nicht vorwärts gebracht wird durch komplizierte Hypothesen, mögen sie auch mit noch soviel Geist und Scharfsinn ausgeklügelt sein, sondern durch einfache Grundgedanken, die an sich evident sind, die aber erst fruchtbar werden, wenn sie zu klarem Bewußtsein gebracht und mit strenger Konsequenz durchgeführt werden".

<sup>2</sup> The discovery of certain prosodic sound variations (in words like *pâte, pâté, pâtisserie*) caused Rousselot, more than 100 years ago, to found the new discipline of Experimental Phonetics devoted to "les modifications phonétiques du langage".

The paper is organized in the following way. We first discuss the relationship between speaking and writing with respect to the word as a phonetic category. Then we try to clarify - in three sections dealing with considerations of principle - the concept of the word as something that can be systematically modified in its phonetic form. In the second part we are going to give three more concrete examples of my basic idea (of systematically modifying the phonetic forms of the words of the given language). Finally we would like to come back to the traditional questions mainly asked by linguists theoretically concerning simplicity and complexity. Concern with simplicity and complexity of speech processes is one thing, another is to consider the interrelations between simplicity and complexity as a phonetic matter of fact when trying to understand the processes of phonetically producing an act of speech. Thus we have the following eight sections:

**Introduction:**

**(1) Speaking and writing**

**Part I: Considerations of principle**

**(2) Sound words and language words**

**(3) Autonymic sound words and heteronymic language words in speech technology (SLP)**

**(4) Random variation and systematic modification**

**Part II: Three examples of practical applications**

**(5) Articulation of complex and elementary sound words**

**(6) CRIL and the central role of the word in the preparation of databases for SLP-technologies**

**(7) PHD: From single words to connected speech**

**Concluding remarks:**

**(8) On simplicity and complexity**

As most of the color pictures and figures presented at the conference in Berlin can not be shown in the printed version, we have prepared a web-version which the reader may find on my home page at the following address:

*<http://www.phonetik.uni-muenchen.de/>*

## **1. Speaking and writing**

As an introduction to what follows I first would like to develop an idea of what - in a narrower phonetic sense - could be called a word (of a given language). We start by looking at the complex relationship that exists between speaking and writing (both involving very complicated, mutually interrelated skilled human actions), and then create a point of departure for the following seven sections.

If we consider what speakers and writers of an utterance are actually producing as pure data, these are so different in nature and form, we would not be able to discover any specific proper-

ties in the written symbols and the recorded speech sounds which could be used to interrelate the two types of data to each other. It is only when we introduce the natural concept of the words of a given language and recognize that we need to know how the words of this language are to be pronounced and how they are to be written that we are in a position to compare any pair of spoken and written phrase of a language. If we recognize that a spoken and written utterance of this language contain exactly the same words in the same order are we inclined to say that they are identical - despite the fact that they are so different in nature and form.<sup>3</sup>

Two written or printed sequences of words such as 'I'm hungry' and 'I'm hungry' are identical if they contain the same sequence of letters including spaces. But it is also true that two phonetically very differently produced utterances of speech are said to be alike (or even identical) if they can be written as the same sequence of words (such as Bloomfield's "I'm hungry" uttered by a child who has eaten and merely wants to put off going to bed or uttered by a needy stranger at the door who wants to express 'please give me something to eat').

In any kind of speech science words are needed to explain the semantic relations that are created by the speaker who expresses himself in an act of speech. These semantic relations exist between the directly observable utterances and what the speaker wants to say (which is not directly observable). And these relations determine what the listener understands and infers when perceiving what the speaker is producing in a given context. This also seems to remain true in a rather technical situation.

When trained listeners look at spontaneously uttered speech data collected for the training of SLP-systems, the production of an orthographic transliteration becomes extremely difficult (if not nearly impossible) as soon as the listener cannot decide which word or sequence of words a given speaker has been trying to utter. In the VERBMOBIL-project, together with our partners in Kiel and Bonn, we have developed tools for handling cases where a word is mispronounced, mutilated, or even unrecognizable. On the other hand, in real speech acts, where we only want to understand the speaker (without having to produce a transliteration of what he is literally saying) we normally ignore these unclear parts of an utterance. Quite automatically we even infer mutilated words from a given context if these words are required for understanding what a speaker is saying. The only necessary condition for any natural speech act seems to be that (i) there is a speaker who creates certain phonetic facts (directly observed by the speaker himself and an audience, if present), and that (ii) the semantic facts (which, with the exception of so-called pairing situations, cannot normally be observed directly) are to be inferred in dependency of observable properties of the speakers utterance, and that (iii) these observable properties can be related to phonetically reproducible words of the language of the speaker.

To provide a point of departure for the following sections I would like to add a new technical detail to my earlier analysis concerning the semantics of phonetic transcription. This small, but helpful detail is simply to distinguish between the different use of two types of quotation marks.

In their analysis of human speech acts logicians and philosophers of language have introduced

---

<sup>3</sup> I have tried to specify the eight main differences between collections of spoken and written language data in Tillmann 1997.

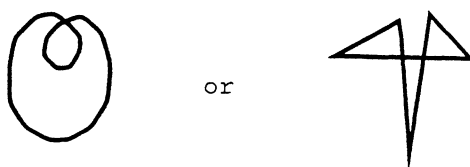
the concept of an *autonymic* use of some categorically established entities. If these entities are written down orthographically quotation marks are employed for indicating an autonymic representation of an utterance. Thus they indicate that a speaker who utters "it is raining" tells the truth iff it is raining. This use of quotation marks as well as the word 'iff' (written with two f-letters to get a shorter version of 'if and only if') may indicate how much philosophers and logicians need writing for their analysis of speech acts! (Which, of course, is true for any kind of speech science; without the invention of writing systems only a few k years ago there would be no science in today's sense at all.)

In my own analysis of phonetic transcription as a way of symbolically representing directly observable phonetic events by graphically well defined entities the starting idea<sup>4</sup> was to show that Gerold Ungeheuer's "extrakommunikative Situationen" may include very specific speech acts in which what the speaker wants to communicate by producing an audible utterance is nothing else but a demonstration of the phonetic form of this utterance. In such a situation the phonetic event is produced by the speaker in an autonymic way, meaning itself.

In speech we don't have the option to employ quotation marks to indicate such an autonymic use of a phonetic event as a reproducible category of its own. Therefore we have to find another solution.

It was here in Berlin that Wolfgang Köhler (before he had to leave Germany) invented and conducted a famous experiment concerning the relation between the phonetic forms of two newly invented meaningless words and their potential psychological meanings. He drew two graphical representations at the blackboard, one with round smooth curves and one with sharp edges and acute angles, and he then asked his students which of these drawings was the meaning of the two words "Maluma" and "Takete", respectively. Not surprisingly, for his subjects "Maluma" sounded round and smooth, whereas "Takete" appeared to be sharp and acute. Thus Wolfgang Köhler got the expected answers.

I used phonetic reproductions of the two pure sound words "maluma" and "takete" to show that (i) new meaningless words can be invented and clearly communicated to an audience under normal noise conditions by just one single demonstrating utterance. Any student is able to give as many equivalent reproductions of "maluma" or "takete" as he is asked for. But I could also use Köhler's examples to introduce the concept of symbolically representing the category of a phonetic utterance. If, in a pairing situation, I pointed to simplified and stylized versions of Köhler's maluma- and takete-signs, i.e.:



I could also prove that (ii) - when in a pairing situation I had pointed to one of these signs and



---

<sup>4</sup> For the first time presented in my inaugural speech on "Phonetik und Sprachliche Kommunikation" at the University of Munich and then further developed in Tillmann with Mansell 1980.



created at the same time just one single corresponding audible "maluma" or "takete" event - any student had understood this as an ostensive definition because he was able to reproduce the sound meaning of the graphically represented sound word.

In Köhler's experiment, the graphical representations were the heteronymic meanings of autonymically introduced sound words. In my modified application the meaning relations are exactly the other way round. I first introduced two signs as such (meaning themselves). To indicate the autonymic use of these signs I propose to use single quotes.


Thus


'' means the sign  (the 'roundsign')

and

'' means the sign  (the 'edgesign')

as a graphically given form or category. (We could even introduce names to identify these autonymically reproducible entities such as the 'roundsign' and the 'edgesign'.) After these signs were introduced as categorically reproducible categories I could - in the pairing situation described above - give them a new *heteronymic* meaning.

To indicate that a graphical representation stands for a phonetically reproducible category I propose to place it between double quotes. Thus the heteronymic meaning of "  " is an autonymically reproducible phonetic event, which (in Tillmann with Mansell 1980) had to be written down in an orthographic form, i.e. "maluma". But we have to keep in mind that the meaning of the expression "maluma" is just one concrete audible reproduction, autonymically demonstrating the category of itself. By only introducing the simple method of making this different use of single and double quotes, we can very easily express that 'ü' means the letter ü, whereas "ü" means a given speech sound produced by a speaker who is demonstrating the category of this phonetic event. Thus when reading in the context of this paper something like "'ü'", just say the soundword "ü".

Without quotation marks, maluma (or Maluma) can have many meanings depending on the context of situation where this word is used. In quotation marks it can have exactly only these two meanings: 'maluma' ≠ "maluma". The first expression means itself (a string of letters: 'm', 'a', 'l', 'u', 'm', 'a'), the second just one concrete demonstration of the phonetic form of "  ".

We may either invent a new writing system for representing complex or elementary sound categories or just use one of the existing ones. In any case it is very helpful to have the distinction

between autonymic and heteronymic uses of sound signs. Thus we can define the more complex expressions

'♡' = 'maluma + takete'

'♡' = 'takete + maluma'

and, as soon as we introduce an ostensive definition and install the heteronymic meaning of '+' by producing "+" or "und", we even know, how "♡" or "♡" is to be reproduced phonetically.

We can also define the sound meanings of single letters such as 'm', 'a', 'l', 'u' by ostensively demonstrating "m", "a", "l" and "u"-events, respectively, and then represent the complexly defined ☺-category by an analysing expression, such as

'☺' = 'maluma'

If the phonetic meaning of 'm', 'a', 'l', 'u' has been effectively demonstrated by the respective autonymic reproductions we can even define quite complex new categories by writing what I call an analysing expression, such as

'☺☺' = 'mumulamulu'

When we are writing we normally are not interested in the letters nor in what they represent phonetically (and, accordingly, we don't have any need to employ single or double quotes). Thus in normal writing situations the user of an alphabetic writing system does not make any references to the written symbols (autonymically), nor to their (heteronymic) speech sound meanings, but to the normal heteronymic interpretations of the words of the language and what they allow the user to express semantically.

## I. Considerations of principle

The history of speech science is a history of discoveries which show how little natural insights human speakers really have in what they are doing when successfully conducting speech acts. Many of the reasons for this lack of insight could be analysed in more detail, but there is in particular one which should be mentioned.

The processes of producing phonetic facts during natural speech acts run much too fast for any direct inspection. Even a trained phonetician looking at the midsagittal representation of a sound sequence such as "ich habe Stolke gesagt" (produced by a speaker of German) has to slow it down by a factor of about 5 in order to be able to observe the relevant details of the jaw and tongue movements. Obviously speech has to be such an automated and highly trained form of behavior and has to run at such high speed to serve the creation of semantic relations that are expressed by these fast actions.

## 2. Language words and sound words

When we compare the phonetic forms of certain words as they are uttered in normal speech acts (where they are used semantically by the speaker in their heteronymic meanings) and as we utter them in isolation in order to autonymically demonstrate, as clearly as necessary, their phonetically reproducible form, we observe all those differences that exist between pure sound words and real language words. Only in the second case do we see forms that are described in pronunciation dictionaries.

In Tillmann with Mansell (1980) we have proposed to specify the second type of phonetically very clear pronounced words as "alphabetically explicit". We also could show how important it is to understand that a phonetic transcription may have only one of two quite different truth value conditions. This depends on what is logically maintained by such a transcription. If the phonetic transcriber is interested in identifying the words of a given utterance he usually separates them by blanks and gives a narrow or broad transcription of their alphabetically explicit forms. In this first case of lexically representing a phonetically given utterance (we used the German term "wörtliche Darstellung" which cannot be literally translated into English) the transcriber is not really interested in specifying the actual phonetic form of this utterance. So the semantics of such a transcription is determined by the claim to maintain that the identifiable words are produced in isolation just as shown in the transcription.

On the other hand we get quite different truth value conditions if the semantics of the transcription is related to the phonetic forms that have actually been produced by a given speaker. As soon as we have to segment and annotate a naturally produced speech signal, the truth values of the resulting transcription are determined according to what can be really observed by a close inspection of the speech wave under auditory control.

It is exactly this kind of relations that exists between the phonetic forms of heteronymically used language words and the forms we observe under the condition of autonymic word demonstrations that we want to move to the center not only of phonetic speech research, but also of phonetic theories of speech acts. Therefore we would like to introduce a new terminology which reflects these two kinds of phonetic forms of words. We will continue to call the words of a language as they are used in their heteronymic meanings in normal speech by the speakers of that language *language words*. But words which are identified by a graphical representation (which can be cited by the use of single quotes) and then are heteronymically interpreted by an autonymic phonetic event (which can be symbolically represented by the use of double quotes) will be strictly called *sound words*. As phonetic objects, sound words can be identified semantically by simply looking at what in earlier publications of the present author has been called their observable articulatory content.

Sound words are always produced in isolation. They can be more or less complex or even quite elementary. Thus "♥" or "♠" are more complex than "☉", "+", or "♠", and "90" is more complex than "9". We are, of course, interested in situations where the lexically given identifier of a language word can be represented by an analysing expression whose components are less complex sound words.

However, if

$$'99' = '9 + 90'$$

we would like to be able to specify, why (and in what phonetic details) "99" is different from "9", "+", "90" produced as a list of unconnected words so that

$$"99" = "9 + 90" \neq "9" \quad "+" \quad "90"$$

becomes true.

We are interested in answering the question how the sound words "9", "+", and "90" have to be changed in their phonetically given (and, to a certain degree, also auditorily observable) articulatory content in order to become a given "99", produced, say, by the same speaker. And then we could look into a database of this speaker to find out how the sound word "99" is in agreement with the phonetic forms of this word when used as a language word. We could also ask the question whether we get at least a slightly different phonetic form if a given sound word such as "99" is just produced as a single language word in a neutral situation without any further context.

Another question is: where do we find sound words in real life? Pure sound words, both elementary and complex ones, are quite naturally produced as spontaneous demonstrations of equivalently reproducible phonetic events in two kinds of situation. When learning to speak a language the words of the language must be ostensibly introduced by a teacher. But also when speakers and listeners of a language start to learn to read and write the words of their own language, sound words play a central role. Mothers produce sound words to demonstrate the category of a reproducible event to their children and teachers produce sound words to instruct their pupils in reading and writing.

Even phoneticians first have to learn to identify the audible qualities of the IPA-sound categories by being exposed to the respective elementary sound words; a good example here would be a proper demonstration of the tense and lax vowels of German in isolation. And here, my personal observation is that the students in our courses learn by themselves to give autonymically presented elementary sound words a heteronymic meaning by referring to their symbolic representations. Thus the vowel of the German word "Kind" is simply identified as the 'small capital i'. So we may have heteronymic and autonymic meaning relations between a sound word and its graphical representation in both directions.

Semantically, the interesting aspect of such cases is that the heteronymic meaning of these alphabetically elementary sound words is just the letter representing the sound word. Here, again, we get Köhler's meaning relation, where the words had a graphical meaning. So the heteronymic meaning of pure sound words would be nothing but their lexical notation such as, for every speaker of German who is not illiterate, the heteronymic meaning of the pure sound word "ü" seems to be primarily the letter 'ü'.

If this observation proves true we could even conclude that for the writers and speakers of a language a written word can take three different possible meanings, while a spoken word has



only two. The German word *Abc*, when written without any quotes, can simply be used as a synonym of the language word *alphabet*; if written as 'Abc' it just means this sequence of letters; and written as "Abc" it has the same meaning as "a-be-tse", pronounced as a sound word by a speaker of Standard German. But the speaker who is producing the word *Abc*, has only two options: the pure sound word means 'Abc', and, when he uses the language word *Abc*, he normally will be referring to the Alphabet.

If there is not a third meaning of a language word (as in the case of *k* = kilo), the single letters in the lexicon such as *a*, *b*, *c*, etc., have obviously only two significant meanings, 'a' and "a", 'b' and "be", 'c' and "tse", etc.

### 3. Autonymic sound words and heteronymic language words in speech technology (SLP)

Our distinction between autonymic sound words and heteronymically used language words is also of interest for some of the actual problems of modern speech technology, in both major domains of SLP<sup>5</sup>, i.e. automatic speech recognition and artificial speech synthesis. Even if pure sound words seem to be of little primary interest in any practical SLP-application, the relations we observe when comparing the phonetic forms of sound and language words may not be ignored, neither in speech recognition nor in speech synthesis. As the latter will be dealt with below in section 7, I restrict myself here to automatic recognition. In this case the relation between sound and language words are determined by the fact that sound words are very clearly articulated, while language words (at least in certain parts of an utterance) are much less clearly articulated. This will have to attract more scientific interest in the context of future SLP-research than it does today. The common aim of phoneticians and speech technologists must be to find methods for deciding whether a given piece of speech is clearly or less clearly articulated by a given speaker, and to what degree this would be the case.

My own career as a professional phonetician started in 1963 with a one year project on automatic word recognition. Together with the brilliant technician of the phonetic institute at Bonn, Herr Rupprath, we created a hardware system consisting of hundreds of transistors, resistances, condensators, etc., with a microphone as input and, as output, 20 small lamps for indicating which one out of 10 Italian cardinals ("zero", "uno", "due", ..., "nove") and 10 additional command words (such as "per", "diviso", "dacapo") had been spoken at each trial.

Fig.1 (in the web-version:) Picture of the first  
DAWID-System, 1964

The original DAWID-System, described at the ICA in Liège (cf. Tillmann et al. 1965), could only be so successful at that time because we made the prudent decision to take whole sound words as the central units to discriminate from each other, and not to try to reduce these com-

---

<sup>5</sup> It should be mentioned that the term SLP (acronym of 'Spoken Language Processing') has been proposed by Hiroje Fujisaki who initiated the ICSLP-Conferences as a common forum for speech science and speech technology. To my understanding phonetics and semantics are going to represent the two major parts of speech science in the SLP-domain.

plex sound words - "zero", "uno", etc. - to more elementary ones (such as "u", "e" or "i", "a", "z", etc). The acronym DAWID, by the way, stands for 'device for automatic word identification by discrimination' (which means that also here in the word DAWID the 'W(ord)' is in a central position).

Our second prudent decision was to concentrate on those acoustic properties of the speech waves whose measurements showed clear maxima for certain speech sounds, so that we could define a threshold for triggering discrete feature detectors. Such properties were, for instance, the frequency of the first formant, F1, the distance of the first and second formants, F2-F1, or the fricative zero-crossing density function. Thus we were even able to use elementary sound words such as "a", "e" or "i", "s" etc. for the testing of single feature detectors and for adjusting their thresholds. However, quite soon it became clear, that the proper triggering thresholds had to be set to a much lower, more sensitive level in order to get the expected feature detections in the case of complex sound words. So we discovered that the measure of certain "distinctive features" of speech sounds depends to a great extent on how clearly the respective sounds are produced by the given speaker in a given sound or even language word.

The probability that there is not a great difference between the phonetic forms of corresponding sound and language words is rather high in the case of isolated word recognition. (In the first DAWID-system we had great problems when we only told our speakers to produce pauses between isolated word productions. Indeed, it is not easy to instruct a speaker of Italian to produce pauses between the production of complex sound words which are longer than the silent intervals of the geminates within the words. In a quite naturally produced Italian "otto otto otto"-sequence the 'tt'-pauses can be about three times as long as the interword pauses.)

During natural speech production the phonetic facts created by the speaker have a clearness-measure that varies between the two extreme H&H-polarities of Lindblom's well-known Hyper-Hypo-dimension. It is not easy to see how this measure could be incorporated into today's HMM-technology. This measure is itself a variable of time, which may change its value from one syllable to the next, depending on factors such as the local tempo of a speech utterance. On the other hand it is quite clear that certain properties of clearly produced sound words may not vanish in any case. If a speaker of German wants to communicate the ownership of some money by uttering either "dies ist mein Geld" or "dies ist dein Geld" the listener must be able to observe the articulatory facts of the "mein"- or the "dein"-soundwords in order to understand which one of these two possessives has been used as a language word. Speech technology still has to conduct considerable research in order to solve the problem how to decide which sound word in the lexicon of a speaker was used by this speaker as a language word in a proper heteronymic function.

#### **4. Random variation vs systematic modification**

Why does phonetic knowledge not play a larger role in modern speech technology? One reason is certainly attributable to the fact, that purely statistical methods such as Hidden Markov Modelling or neural net computing produce much better results than so-called rule based expert systems. The situation is somehow self-contradictory, without any recognizable ways of effectively combining statistical and knowledge based methods. On the one hand rule based systems are simply much too powerful in two respects: they generate exploding sets of possi-

ble solutions that cannot be effectively computed in a reasonable time, and these sets also remain more or less empirically empty; and, theoretically derived forms are of little practical interest if there is not one real single utterance in a given database that falls under the specified category and could be taken as an instantiation of it. On the other hand, we must consider that even the largest databases of spoken utterances that have been collected up to now for the purpose of training and evaluating SLP-systems do not contain enough material to model most theoretically interesting cases. This is the dilemma of modern speech science.

The research aim in this situation can only be to find ways to treat phonetic variability (and to reduce purely statistical variation) by introducing the concept of systematic modification. A first step of reducing the amount of variability could simply consist in separating the data of individual speakers. Another way of reducing acoustic variability could consist of relating the acoustic picture of sound and language words to the underlying articulatory processes which produce that picture, and then interpret the individual articulatory data by relating it to a generalized system. This could one day be obtained by means of a properly organized neural net which turns heard speech signals in some generalized newly articulated speech waves. A second step will be to look more closely at the prosodic form of speech productions analysing glottally controlled voice production and segmentally controlled sound articulation as a whole, i.e. in a totally intergrated way.<sup>6</sup> But the most important step would be to specify each lexically represented sound word with respect to its possible modifications that quite systematically have to take place as soon as this word, in a specified context of other words and of situation, is to be uttered by a given speaker as a real language word. I would like to illustrate this idea by some observations in Barbara Kühnert's dissertation, where she analyzed the t\_k-assimilation of native speakers of German and English.

The German word "Blatt", produced as a sound word, shows a very clear final t-articulation. Used as a language word in a context like "Das Blatt kam von der Eiche" the t-behaviour of the tongue tip depends on whether this sequence is produced in normal or in fast speech. In very clear speech the language word still has a measurable EPG-t-contact; if this contact is lost, the electromagnetically measured movements of the front tongue may still show a whole scale of reductions. Barbara found in her data everything, (i) the tongue tip still contacting the alveolar region, (ii) the tongue tip moving almost all of the way up without getting into EPG-contact, (iii) the tongue moving half of the way, (iv) only a little bit, and (v) not at all. Thus there obviously is an H&H-continuum, that determines the behaviour of the tongue tip at the end of words with a final t-sound followed by a k-word.

That this articulatory reduction of the t-sound in a given situation of t\_k-assimilation is not just random variability, but a very systematic modification depending on the situation, seems quite clear to me. In the final section of her dissertation, '7.5 Der Sprecher als Hörer' (p. 354 in FIPKM 34) Barbara Kühnert describes an experiment with two subjects (an English and a German phonetician) who had to judge their own reduced t\_k-productions. The stimuli were

---

<sup>6</sup> I still have some hope that my early proposal to handle what I called "silbischer Ausprägungs-kode" (meaning the locally varying degree of 'well-articulatedness' of alphabetically specifiable soundsequences, the local degree of clearness of alphabetic soundstructure, the local tempo of syllable production, etc.) could consist in defining it as a computable function of intonation. Campbells prosodic concatenation system CHATR seems to offer a first verification of this basic idea.

short VCV-segments cut out from the respective reduced t\_k- as well as from k\_k-productions as control items.

The German subject (which was me) did quite well with his own reduced t\_k-productions. My hypothesis is that in my productions there is - even if the t-movement of the tongue tip is reduced to zero - still a prosodic reflex of reduced segmental information that allowed me to reach a 100 % score in comparison with the control sequences. But this is a hypothesis that needs further investigation.

My hopeful conclusion concerning the possibility to separate the variability within the phonetic forms of language words into the components of statistical variation and systematic modification is strongly supported by this result. Only if in the two fast t\_k- and k\_k-versions there is still some systematic prosodic reflex of the missing t-movement the outcome of the experiment could not be explained as an artefact of something else.

## **II. Practical Applications**

That sound words give us just the material we not only need, but can also simply take from any speaker of a language if we want to build up new utterances (of the same kind we observe in natural speech acts and which will be accepted by the speakers and listeners of that language as quite naturally produced utterances), is indeed a very challenging idea which I would like to further illustrate by describing the following three examples I borrow from ongoing projects in our institute.

### **5. The articulation of complex and elementary sound words**

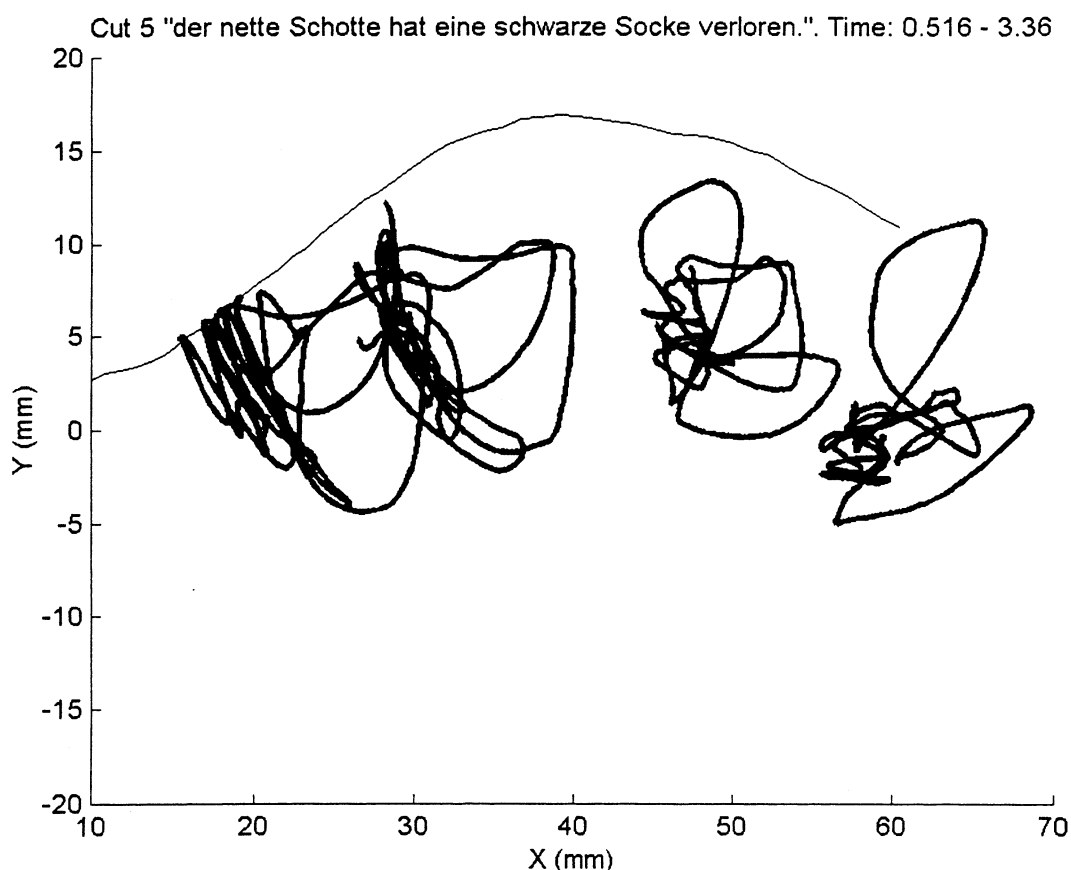
Probably the best way to understand which problems have to be solved in trying to approach our challenging goal is to look at some of our articulatory recordings. The data has been collected with our electromagnetic equipment (cf. Hoole 1996) which allows us to record up to ten fleshpoints from the speakers lips, jaw, and from about 5 cm of the front part of his or her tongue during normal and fast speech productions.

First of all, I should however mention that none of these articulatory projects is explicitly devoted to the ambitious goal which I'm talking about here. In none of our applications for receiving our project-grants in articulatory phonetics has the role of the sound word as a central phonetic unit actually been mentioned. Thus it should be clear that in our research we are still dealing with much more specific questions that can be answered by analysing the data itself (without trying to modify them in a proper way).<sup>7</sup> The following picture illustrates the movement of the front part of the tongue during a normal and a fast production of the utterance

---

<sup>7</sup> On this occasion I should specially thank the German Research Council DFG for sustaining our work by grants Ti 69/29 (articulation of the German vowel system), /30 (development of 3-d-EMA, with only one 'M'), /31 (our contribution to the 'DFG-Schwerpunkt Sprachproduktion'). The work in these projects has been or is done by Phil Hoole, Barbara Kühnert, Andreas Zierdt, Christian Kroos, Christine Mooshammer, and Anja Geumann.

"Der nette Schotte hat eine schwarze Socke verloren" (produced by three speakers).



**Fig. 2** (cf. the color-versions of this example on the web)

Without listening to the acoustic results of these speech movements no trained phonetician will be able to decide which complex or elementary sound words have been transformed here into the language words by the three speakers who were reading the prompting word sequence in two different speeds.

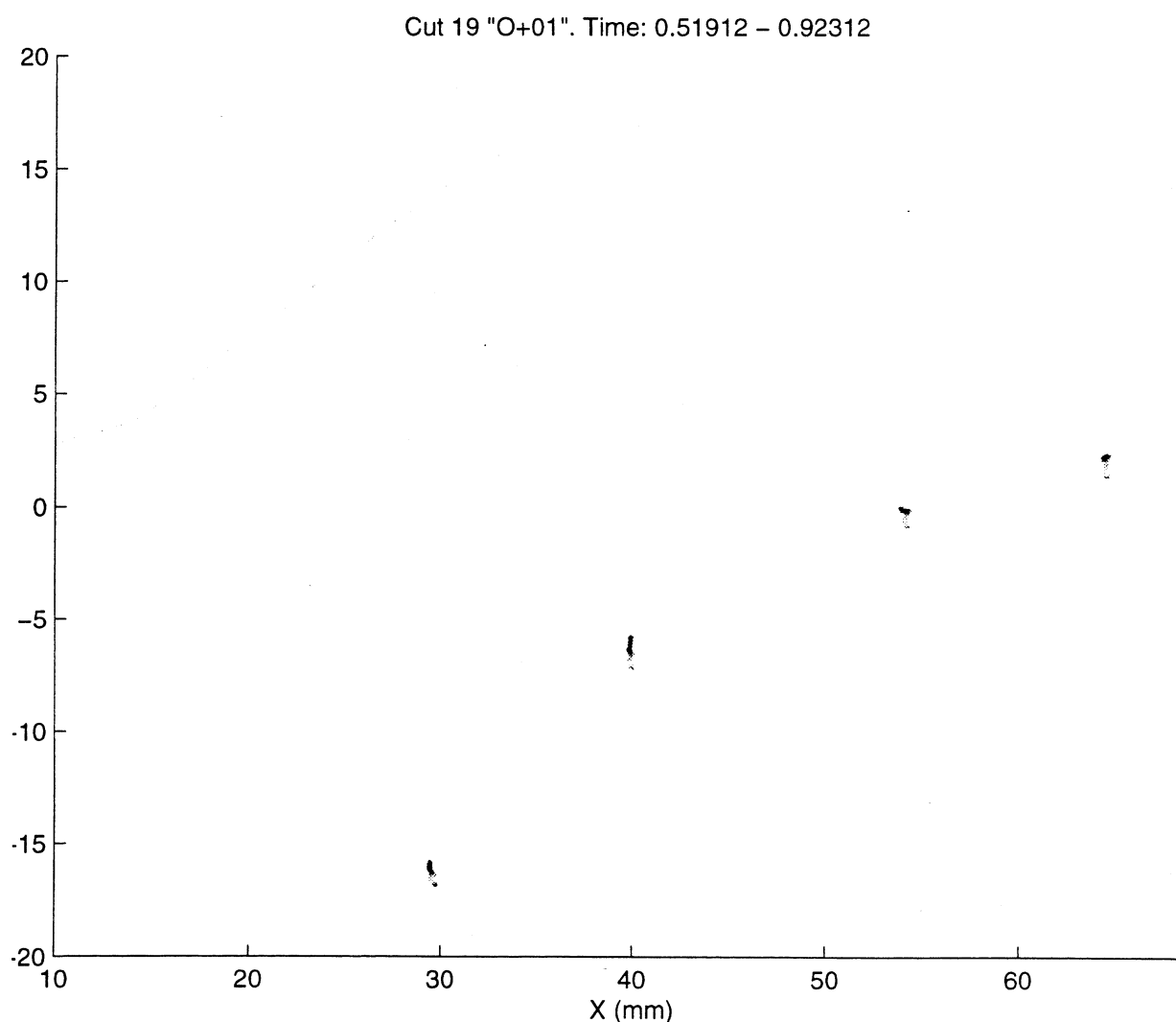
To get a better insight into the obviously very complex processes of transforming isolated sound words into a sequence of real language words I would like to propose to begin with elementary sound words such as "m", "a", "l" etc. and then to see, how these have to be modified in order to form a lexically given "maluma".

As everybody knows, things are a little bit more complicated than this would imply. We only have to mention the most elementary processes of coarticulation and assimilation at the lowest level to be reminded of this. A good example is the mid vowel tongue position of the tense and lax German vowels in dependency of the CVC-context as in the case of  $C = p/t/k$ .

On the other hand we have to consider that certain alphabetically represented sound words are not really elementary. Sounds such as [k], [g], [t], [d], or [p] and [b] cannot be demonstrated as elementary sound words without also producing a voiced or voiceless vowel.

A much more complicated "elementary" situation is given in the case of the so-called German

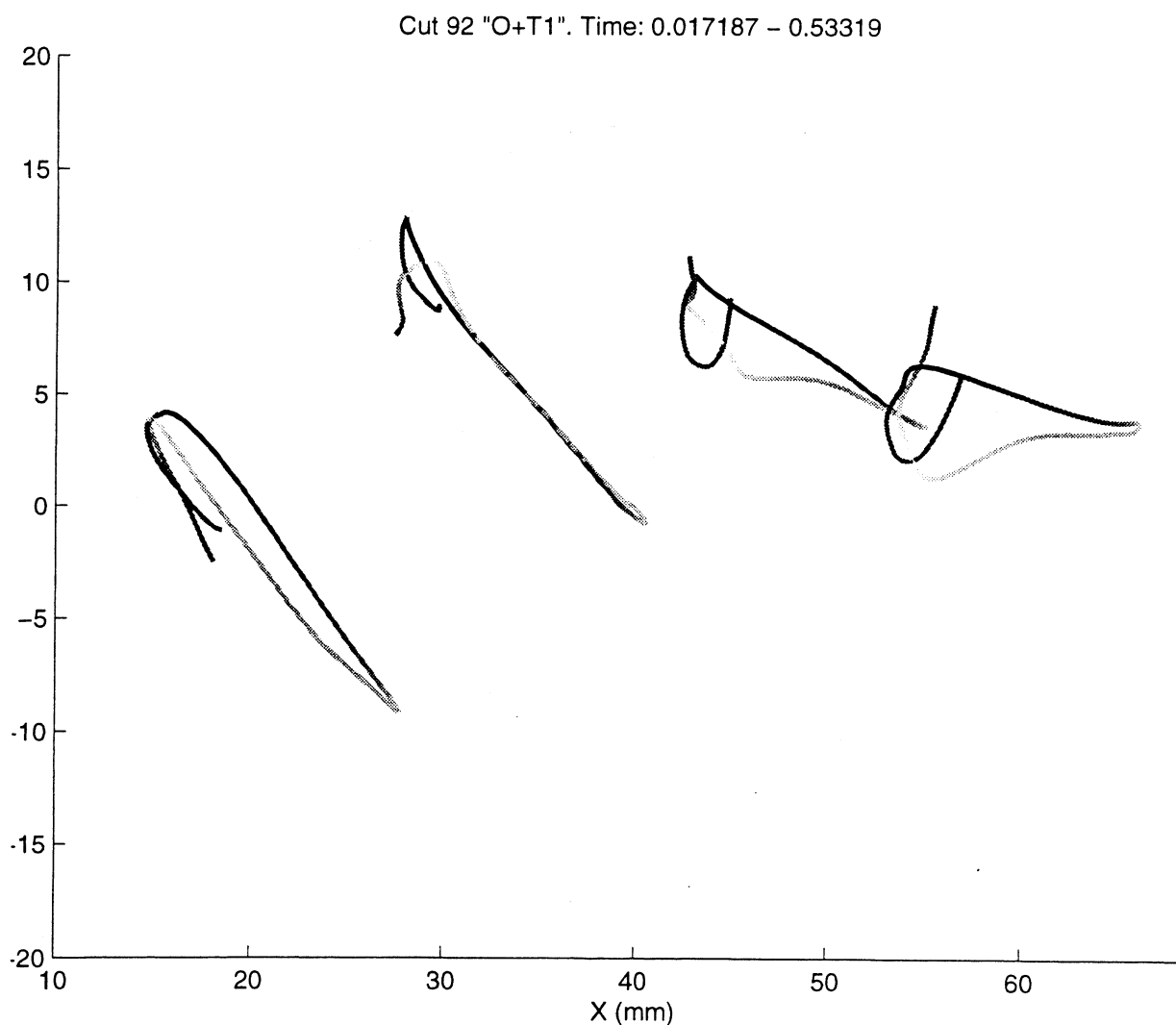
vowel opposition, which is observed in pairs such as "Miete, Mitte", "Mühle, Müller". No phonetically untrained speaker of German is able to demonstrate the vowel of "Mitte" or "Müller" as an isolated single-vowel sound word. The Standard German sound system contains only those elementary sound words which are represented in the lexicon as "i", "e", "ä", "ü", "ö", "a", "o", and "u" which are listed in the Duden for instance by their respective letters. Here, by the way, I would like to confess that (unlike most of my colleagues today) I totally agree, as a phonetician, with Theo Vennemann's analysis, which clearly shows that from a phonological point of view the German vowel opposition can not be reduced to an elementary sound word opposition. The distinction is certainly a prosodic one and can only be dealt with in closest connexion to the syllable structure of German<sup>8</sup>. I would even propose to introduce the concept of prosodic modification in its strongest sense. So we simply take, for instance, a generalized articulation of demonstrating an elementary German "o"-word



**Fig. 3** (cf. the color-versions of this example on the web)

<sup>8</sup> See also T. Becker 1998 and D. Restle 1998

not only for modifying it into a complex sound word such as "gepope", "getote", "gekoke" (in two different tempos):



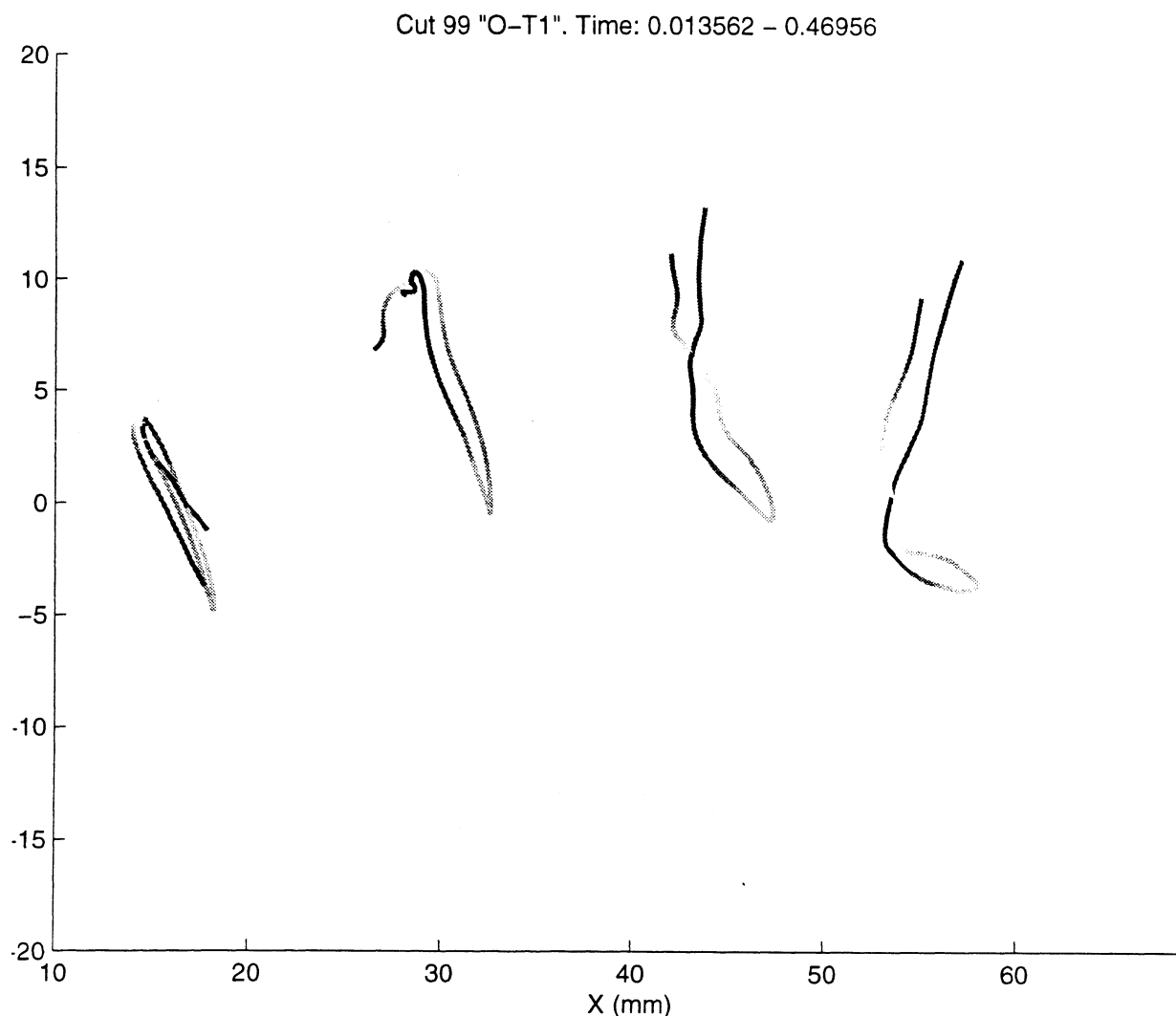
**Fig. 4** (cf. the color-versions of this example on the web)

but also to systematically change it into the "gepoppe", "getotte", and "gekocke" sound words (also in the respective tempo versions).

(see Fig. 5 on next page)

Everybody can see which kind of transformations have to be taken into account in order to achieve our ambitious aim.

When we have arrived at a theory which specifies all necessary algorithms for the linear and nonlinear interpolation between elementary sound words and their complex counterparts in the sound word lexicon of a speaker, we shall also be in the position to do the next step of developing the theory that further transforms complex sound words into proper language words (so that they look just as those in *Fig. 2*, above).



**Fig. 5** (cf. the color-versions of this example on the web)

The details of such an articulatory theory of producing natural utterances for conducting real speech acts will be extremely complicated (even if most of the work could probably be done by neural net computing methods), but the basic idea remains quite simple, that is of taking the sound words of a language as the central units. We are convinced that in the case of speech synthesis it is much easier to take something already given and change it into something more or less different than to create something from nothing.

The words to be changed must not only be stored as data, but also categorically specified within the lexicon of either a given individual or a computed generalized speaker of the language in such a way that all possible modifications of the given sound words as well as all necessary transformations for producing proper language words are sufficiently determined. Here particularly phonologists can do a very helpful job if we phoneticians supply them with an



experimental artificial speaker.<sup>9</sup> For instance, the sound words of Standard German must be specified with respect to which syllables are "tone-syllables" in the sense of Thomas Becker and so must be prosodically processed with respect to the vowel opposition of German, and which are not and are to be treated differently.

## **6. The central role of the word in the preparation of databases for SLP-technologies**

The second example should only be mentioned here, because Christoph Draxler is giving a whole talk just on this topic. Let me shortly address some of the essential viewpoints.

First of all I want to point out that in PHONDAT (and later in VERBMOBIL) we had a close collaboration between the phonetic institutes of Kiel, Bonn, and Munich, and especially Klaus Kohler and the present author took great care that the CRIL-conventions of the IPA were strictly applied in the organisation of all our speech data collection. In this context we introduced the term 'canonic word form' which to my personal understanding could also be used to refer to a clear lexical representation of the corresponding sound word as described in a narrow SAMPA-notation. Christoph Draxler will come back to this term because he has translated it into a PROLOG-predicate which plays the central role as a unit for organizing the databases of PHONDAT (and VERBMOBIL).

CRIL is the acronym of 'computer representation of individual languages'. The conventions have been defined at the Kiel convention of the IPA and say that there should be at least three levels of symbolic annotations for any given speech signal: (i) an orthographic representation (if possible) to guarantee the identification of the lexical entities produced as language words within the given utterance; (ii) a broad phonemically oriented notation of a possible citation form, which corresponds to our complex sound words; (iii) a narrow transcription of what has actually been pronounced by the given speaker which corresponds to what I refer to as the phonetic form of the actually produced language word. This third level is directly related to the speech signal and indirectly related to the canonic forms by systematically indicating insertions, substitutions, and deletions of sound segments. What is still missing (and has not yet found a regulation with respect to some CRIL-convention, but will be certainly needed in the near future) is a method of specifying the degree of articulatory clearness ('Wohlartikulierteit'). My prediction is that these prosodic components of a given utterance will be handled in the near future by some automatically derivable measures which are directly related to the phonetic properties of the speech signal at non-symbolic level.

Using speech data (which had been labeled in Kiel and Munich according to the CRIL-conventions) we have developed the MAUS-system that automatically segments and labels spontaneously produced speech utterances under the condition that (i) the sequence of language words within this utterance is orthographically specified, and that (ii) a canonic representation of the corresponding sound words can be looked up in some kind of a pronunciation dictionary containing a regularly defined canonic word form.

In this context, any word of a given language (or the ideolects of the speakers of this language) could be seen as a theoretically specifiable object that contains all possible systematic modifications in relation to its canonic sound form (as can be seen in the graphs of Florian Schiel's

---

<sup>9</sup> See also Kohler 1991.

contribution to this workshop). The MAUS-system works only under the condition that the canonic form is known and we can specify a corresponding theoretical object containing all possible segmental modifications of the sound word when used as a language word. In the analyzed speech wave, these can be empirically verified with respect to the actually given form of the language word as it was uttered by the speaker. In extreme cases, as we all know, the phonetic form of a sound word when used as a language word can be reduced to zero.

Florian Schiel will describe the MAUS-system in his contribution to this workshop in more detail. I may restrict myself to say only two more things. I wish to confess that I am proud, indeed, about the fact that our idea of verifying the phonetic forms of language words given only their descriptions as sound words proved to work so beautifully, even in its present state (which will be further developed of course, in the near future). Secondly, I wish to emphasize again that it is only because we took the word as a central phonetic unit we were in a position where we could start to automatically collect the knowledge we need for developing a CPT of German<sup>10</sup>.

### **3. PHD: From sound words to connected speech**

My last example is almost future music, because many components of this new project are still in planning stages. The acronym PHD stands for 'parametric high definition speech synthesis'. Together with Hartmut Pfitzinger and Kurt Kotten I have begun to develop an experimental system that will allow us to systematically modify sound words produced by a given speaker and transform these then into language words acoustically.

Several years ago, in a DFG-Schwerpunkt (Sprachpsychologie) we designed and realized a package of DSP-programs to define a continuum between different speakers uttering the same sequences of language words. The system interpolated between presegmented parts of these utterances and produced a set of stimuli for conducting experiments on categorical perception of speaker identities (cf. Tillmann et al. 1984). In the new PHD-project we are, at least at the beginning, less interested in interpolating (and extrapolating) inter-individually between 'the same utterances' of different speakers, but between different sound forms of the same words produced by the same speaker. We are developing these methods of intra-individual interpolation between given utterances and thereby producing acoustically new utterances because we believe that an experimental work bench of this type is needed to learn more about what it means that an utterance of a word can possess a variable degree of clarity. How do we have to reduce phonetic properties and time durations of a given sound word to transform it into a realistic language word of the particular speaker in the proper contexts? Thus, quite differently from our ideas in the articulatory oriented example above, in the PHD-project we are less interested in combining elementary sound words to the complex sound words, but in reducing complex sound words into realistic language words.

The PHD-system is not designed as a typical text-to-speech system, but (in a certain sense) as a 'language\_word\_system', and we are designing this system mainly for conducting experimental investigations concerning the central question that I think is the most important one

---

<sup>10</sup> The goal of developing a strictly empirically based 'complete phonetic theory' of a spoken language has been proposed by Pompino-Maschall and the present author in our contribution to the EUROSPEECH conference in Berlin, 1993.

that phonetic speech research has to answer in the near future: How can we theoretically and also practically relate the phonetic forms of language words to the clearly defined properties of the corresponding sound words as they are autonymically demonstrated by the speaker.

What we do have to change when going from one speaker to another speaker of the same language is quite another question that can only be answered as soon as we know what these speakers already do by themselves when they are changing their data intra-individually from one language word to another one and in relation to the given sound word.

## **8. Concluding remarks on simplicity and complexity**

One of the traditional goals of linguistics has always been to reduce the complexity of phonetically given utterances to a simple underlying grammatical representation. The only way, I think, I could agree which this traditional goal would be to reduce any regularly produced speech utterances to the words which are contained in those utterances, and representing these words for any given utterance by a lexically specified object which is the set of all its possible modifications under defined conditions. Such a theory will be a rather complex one, but it will be governed by the clear and simple idea that the word is the central unit of speech production and speech perception.

There is still quite another aspect concerning the complexity or simplicity of speech processes. All phonetic facts which are to be theoretically modelled by a phonetic object incorporating a complete picture of the potential segmental and prosodic sound structures of each single word of a spoken language, are facts given at the periphery of the speaking nervous systems. The speakers and listeners are able to - and have to be able to - directly observe them as they appear during each act of speech. If we call these facts in comparison with their linguistic descriptions complex, we must, on the other hand, understand that they are (relatively) extremely simple - if we only compare them with those (really extremely) complex processes that have to take place within our human nervous systems during any act of speech, be it an act of demonstrating an elementary sound word, say "o", or conducting a real act of speech.

If we only look at the articulatory content of sound and language words I believe that, in the near future, we will begin to understand that the particular phonetic form which a word takes in a given utterance, is nothing else but a computable prosodic function of what the speaker wants to express semantically.

## **References**

- Becker, T.: Das Vokalsystem der deutschen Standardsprache. Frankfurt 1998
- Draxler, Ch.: Database systems for spoken language corpora. (this volume)
- Hoole, P.: Theoretische und methodische Grundlagen der Artikulationsanalyse in der experimentellen Phonetik. Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München (FIPKM) 34, 3-174, 1996
- Kohler, K.: Prosody in speech synthesis: the interplay between basic research and TTS application. *Journal of Phonetics* 19, 121-38, 1991

- Kühnert, B.: Die alveolare-velare Assimilation bei Sprechern des Deutschen und Englischen: Kinematische und perzeptive Grundlagen. FIPKM 34, 175-392, 1996
- Paul, H.: Prinzipien der Sprachgeschichte. 1898. Zit. nach: 5. Aufl., Tübingen 1920
- Restle, D.: Silbenschnitt - Quantität - Kopplung. Phil.Diss., München 1998
- Schiel, F., and Kipp, A.: Probabilistic analysis of pronunciation with 'MAUS'.  
(this volume)
- Tillmann, H. G.: Das phonetische Silbenproblem. Phil. Diss., Bonn 1964
- Tillmann, H. G.: Eight main differences between collections of written and spoken language data. FIPKM 35, 139-144, 1997
- Tillmann, H. G., Heike, G., Schnelle, H., Ungeheuer, G.: DAWID - Ein Beitrag zur automatischen Spracherkennung. Paper A 12, 5th ICA, Liège 1965
- Tillmann, H. G., mit Mansell, P.: Phonetik. Stuttgart 1980
- Tillmann, H. G., Schiefer, L., and Pompino-Marschall, B.: Categorical perception of speaker identity. Proc. 10th ICPhS, 443-449, 1984
- Tillmann, H. G., and Pompino-Marschall, B.: Theoretical principles concerning segmentation, labelling, and levels of categorical annotation for spoken language database systems. EUROSPEECH'93, 1691-1694, Berlin 1993